Homework 3 **CSCE 421**
Due: 7.00 pm on October 21, 2019

Instructions for homework submission
a) For the **math problems**, please typewrite your answers in Latex, or handwrite your solution *very clearly*.
b) For the **experimental problems**, please write a brief report. At the end of the report, please include your code. Print the report, including the code.
c) **Staple all your answers and hand them out in paper in class or in the instructor's office.**
d) Please start early :)
e) The maximum grade for this homework, excluding bonus questions, is **10 points** (out of 100 total for the class).

Question 1 (6 points)
Predicting forest fires: Forest fires are a major environmental issue endangering human lives. This renders their fast detection a key element for controlling them and potentially preventing them. Since it is hard for humans to monitor all forests, we can use automatic tools based on local sensors to do that. Through these sensors we can get information regarding the meteorological conditions, such as temperature, wind, relative humidity (RH), and amount of rain. We can also compute several fire hazard indexes, such as the forest fire weather index (FWI), fine fuel moisture code (FFMC), duff moisture code (DMC), drought code (DC), and initial spread index (ISI). Using these measures, we can predict whether fire is going to occur in the forest, as well as to estimate the amount of burned area. Such data are part of the "Forest Fires Data Set" of the UCI Machine Learning Repository and their description can be found here: `http://archive.ics.uci.edu/ml/datasets/Forest+Fires`.

Inside "Homework 3" folder on Piazza you can find one file including the data (named "hw3_question1.csv"). The rows of the file refer to the data samples, while the columns denote the features (columns 1-12) and the outcome variable (column 13), as describe bellow:

1. **X**: x-axis spatial coordinate of the forest: 1 to 9
2. **Y**: y-axis spatial coordinate of the forest: 2 to 9
3. **month**: month of the year: 1 to 12 to denote "jan" to "dec"
4. **day**: day of the week: 1 to 7 to denote "mon" to "sun"
5. **FFMC**: FFMC index from the FWI system
6. **DMC**: DMC index from the FWI system
7. **DC**: DC index from the FWI system
8. **ISI**: ISI index from the FWI system
9. **temp**: temperature in Celsius degrees
10. **RH**: relative humidity
11. **wind**: wind speed in km/h

12. **rain**: outside rain in mm/m2
13. **area**: the burned area of the forest (this is the **outcome** variable)

(a) (1 point) Data exploration: Plot a histogram of the outcome variable (column 13). What do you observe?

(b) (5 points) Classification: Based on our observations from the previous question, we can dichotomize the outcome variable, based on whether its corresponding value is zero or greater than zero. This creates the following two classes:
Class 0: Forests not affected by the fire, i.e. area = 0
Class 1: Forests affected by the fire, i.e. area > 0
After dichotomizing the outcome variable, we can run a classification task to *predict whether or not fire will occur in a certain forest* based on the input features.
Use a logistic regression classifier to perform the above binary classification task using a 10-fold cross-validation. According to this, we randomly segment the data into 10 sets. Each time one sets acts as the testing, while the rest of the data are the training data. A logistic regression classifier is trained on the training set and tested on the testing set. This process is repeated 10 times and we report the average accuracy over the testing sets from each fold.
Note: You do not need to implement the logistic regression, you can use existing libraries. You need to implement the 10-fold cross-validation.

Question 2 (4 points)
Maximum Likelihood Estimation: The goal of this problem is to model the number of tickets sold in a train station between August 25th, 2012 to September 25th, 2014.

Inside "Homework 3" folder on Piazza you can find the file including the corresponding data (named "hw3_ question2.csv"). Column 1 refers to the day/time of the year, and Column 2 refers to the number of tickets sold during that corresponding hour. The file includes around 18K columns, which correspond to the hours of the data.

(i) (0.5 points) Data exploration: Plot the number of tickets sold over these two years. The x-axis should represent the hours elapsed since August 25th, 2012. The y-axis should represent the number of tickets sold over each hour. What do you observe?

(ii) Statistical model of first month of data: For this question we will only consider the data that correspond to the **first month** of the study (i.e., August 25th, 2012 to September 24th, 2012). Let $\{x_1, \ldots, x_N\}$ be the the hourly number of tickets of the first month, where in our case $N = 744$.

(ii.a) (0.5 points) Plot the number of tickets sold over this first month. The x-axis should represent the hours elapsed since August 25th, 2012. The y-axis should represent the number of tickets sold over each hour $\{x_1, \ldots, x_N\}$. What do you observe now?

(ii.b) (1 point) We can model the first month of data using a Poisson distribution. Poisson distribution is ideal for this task, as it expresses the probability of a given number of events occurring in a fixed time interval. Therefore we can express the probability of selling x number of

tickets within each hour of the first month as a Poisson distribution using the following equation $f(x) = \frac{e^{-\lambda x}\lambda^x}{x!}$. The parameter λ is called the rate of the Poisson distribution and this is what we would like to estimate from our data. Assuming that the hourly number of tickets of the first month $\{x_1, \ldots, x_N\}$ are independent, find the mathematical expression of the likelihood $l(\lambda)$ of the data for the first month.
Hint: You have to compute the product $l(\lambda) = \prod_{n=1}^{N} f(x_n)$ and substitute $f(x_n)$ with the given expression of the Poisson distribution.

(ii.c) (0.5 points) Find the mathematical expression of the negative log-likelihood $L(\lambda)$ of the data from the first month assuming the above Poisson distribution.
Hint: You have to compute $L(\lambda) = -\log l(\lambda)$.

(ii.d) (1 point) Find the maximum likelihood estimation of λ, i.e., find the mathematical expression of the minimum of the above negative log-likelihood $L(\lambda)$ with respect to λ. What do you observe?
Hint: You have to compute the first order derivative of $L(\lambda) = -\log l(\lambda)$ with respect to λ and set it to zero.

(ii.e) (0.5 points) Based on the above mathematical expression, find the maximum likelihood estimation of λ from the given data $\{x_1, \ldots, x_N\}$.

(iii) (Bonus) Statistical model of all the data: For this question we will consider all the data samples, which will be split into train and test (as described below).

(Bonus - 1 point) (iii.a) Let $\{x_1, \ldots, x_M\}$ be the the hourly number of hourly tickets between August 25th, 2012 to May 24th, 2014, where in our case $M = 15312$. We will now assume that the rate of the Poisson distribution is not constant, but depends on the hours t_n elapsed since August 25th, 2012, i.e., $\lambda_m = e^{\theta_0 + \theta_1 t_m}$. Find the mathematical expression of the likelihood $l(\lambda)$ and the negative log-likelihood $L(\lambda)$ of the data $\{x_1, \ldots, x_M\}$. What do you observe? Can this expression be solved analytically to find the minimum?

(Bonus - 1 point) (iii.b) Assuming that $\theta_0 = 450$, find the maximum likelihood estimate of θ_1 using batch gradient decent.
Hint: You have to find the minimum of the above negative log-likelihood $L(\lambda)$ with respect to θ_1 using gradient descent, therefore you will compute the first order derivative of $L(\lambda)$ with respect to θ_1.

(Bonus - 0.5 points) (iii.c) Assuming that $\theta_0 = 450$, use mini-batch gradient decent for the above problem to find the maximum likelihood estimate of θ_1.

(Bonus - 0.5 points) (iii.d) Using the model with the estimated parameters from gradient descent, predict the number of tickets that will be sold for each hour between May 25th, 2014 to September 25th, 2014. Compute the residual sum of squares error between the actual and predicted values.